图书情报工作
第64卷 第8期 2020年4月

基于语义和位置相似的作者共被引分析方法及效果实证

■ 张汝昊
中国科学院成都文献情报中心 成都 610041 中国科学院大学经济与管理学院图书情报与档案管理系 北京 100049
中国科学院文献情报中心 北京 100190

摘 要:[目的/意义]作者共被引分析是探索领域知识结构的重要方法,在复杂的学科发展态势下,其依赖于共被引频次的作者关联度度量颇具争议。对此,提出一种基于语义和位置相似的作者共被引分析改良方法。[方法/过程]在介绍基本原理的基础上,以图情领域为例开展基于语义和位置相似的作者共被引分析改良方法的效果实证,面向CNKI期刊库进行引文全文挖掘,并对引用句及引用位置进行抽取,结合预训练的领域词嵌入模型计算共被引文献间的深层相似度和作者间的关联强度,利用网络分析和因子分析法对比该方法与传统方法的效果差异。[结果/结论]结果证明,基于语义和位置相似的作者共被引分析改良方法能更准确地识别共被引作者的关联强度,可发现更为细致的学科知识结构,并具有可拓展性与可应用性。

关键词:作者共被引分析 引文内容分析 共引位置分析 全文本引文分析 领域知识结构
分类号:G250
DOI:10.13266/j.issn.0252-3116.2020.08.013

1 引言

作者共被引分析(author co-citation analysis, ACA)是文献研究的主要分析方法之一,由H. D. White和B. C. Griffith于1981年首次提出[1]。其基本假设是:当作者A与作者B同被文章C所引用时,则A与B之间具共被引联系R,若A与B同时被引用的次数越多,则联系R就越为密切。在这一方法假设下,所有共被引的作者间将具有某种主题上的关联,并在一定范围内呈现出明显的主题聚类分布。因此,ACA常用于学科知识结构的识别和科学共同体的发现[2]。

伴随着近年来科学研究的推进和技术的不断革新,各学科正快速交叉融汇,新的研究领域不断以难以探测的方式悄然诞生,学者所从事的研究主题也愈发多元,这都对传统ACA方法提出了挑战。传统方法仅依赖于有限的著录信息,将基于共现与否(非1即0)的次数统计作为作者间联系强度的依据,这是朴素的、很具争议的[3],其忽略了共被引作者间的真实而深层的联系,准确性、科学性都难以保证。

笔者提出了一种基于内容语义和位置临近的ACA方法(semantic and proximity-based author co-citation analysis, SPACA),以共被引作者所在引用句的内容和出现位置的相似性来度量作者间的关系强度,以克服传统方法的缺点。通过采集中国知网(CNKI)期刊数据库图情领域内的文献全文文本,开展该方法与传统共被引分析方法ACA的对比实证,探索这一改良方法的特点与可应用性,以推动ACA方法的精准化革新。

2 相关研究回顾

近年来,自动化文本处理技术日趋成熟,文献全文文本已逐步能以半结构化形式在数据库(如PubMed、BioMed Central、Citeseer、arXiv等)中获取,这些条件使得基于全文文本的引文内容分析应运而生并愈发受到重视。这是一种能深入施引文献全文,获取引用强度、引用位置、引用功能、引用语义等微观的引文全文内容数据,以量化计算引用所代表的关系强度、影响程度等新型引文分析方法。当前国内外引文内容分析法研究可分两个层次[4-6]:包含引用主题分析[7-9]、引用功能分析[10-12]、引用范围识别[13-15]在内的关注引用内容本身的引用语义层次,以及包含引用位置分析[16-19]、

作者简介:张汝昊(ORCID:0000-0002-4372-8726),硕士研究生,E-mail:zhangruhao@mail.las.ac.cn。
收稿日期:2019-09-27 修回日期:2019-12-17 本文起止页码:111-124 本文责任编辑:徐健

引用强度分析[20-21]等在内的关注引用内容外在特征的引用语法层次。

引文内容分析为传统的共被引理论由面向著录信息深入全文文本信息提供了契机,许多学者尝试对其进行改进。其中,一些学者关注于利用全文文本中的引用位置接近程度代替原有的共被引计数。如 A. Elkiss、A. Callahan、S. Liu 与 C. Chen、B. Gipp 与 J. Beel、J. An 等均以不同的方式证明了文档中共被引文献之间的实际相似性与它们在文本中的接近程度有关,并提出利用引用临近索引(Citation Proximity Index, CPI)、引用章节相似性等位置临近性测度方法计算作者联系强度[22-27], M. Eto[28]、赵蓉英[29]证明了在同被引分析中考虑同被引位置的影响,能够提升检索效果和共被引聚类的效果,有利于进行深层次研究和评价。另一些学者则关注引用句本身,以词频、TF-IDF、LDA、C-Value 等算法抽取引用句的特征词或主题进行内容表征和相似性计算,如 Y. K. Jeong 等[30]、K. Lu 与 D. Wolfram[31]、祝清松[32]。随着研究的不断深入,一些学者提出了综合性的共被引分析改良方法,其中,刘盛博将共被引句根据位置邻近度划分为多个级别,通过计算各级别内共被引句的内容相似度,探寻位置临近对共被引内容相关度的影响程度,从而更科学地确定引用位置权重取值[33]; H. J. Kim 等则利用章节位置构建细粒度的作者共被引矩阵,对引用句进行词频相似度计算,在 PubMed 数据库中开展肿瘤领域的实证研究,取得较优效果[34]。此外,受全文数据获取困难的影响,国内学者引入内容文本信息对共被引分析的改进研究仍以利用著录信息为主[35-37]。

正如 Y. Ding[38]和赵蓉英[39]所指出,当前研究对全文文本的利用仍不够深入,实证研究在国内尤为缺乏。目前相关研究关注点大都单一,而对于引用主题的分析多停留于引用句本身的固定词项频率、句法结构、小范围概率模型,而少有深入语义层面,使得相关的共被引文献间的联系仍然是浅层的、基于语句外在特征的。此外,相关研究较少从作者层面出发,相较于文献级,作者共被引分析更具后期应用价值,但作者相较于文献也更为"多面",依赖引用次数的关系强度易造成作者间联系揭示不准确、聚类结果不理想的问题,因此,充分利用全文文本信息对作者共被引分析进行改良具有现实意义。

3 基于语义和位置相似的作者共被引分析法设计

笔者提出的改良型方法 SPACA 充分利用了共被引文献所在引用句的内容语义信息和所在章节位置信息,在此基础上进行相似度计算,并以作者文献集中的最大相似度值作为作者间的相关强度值,以此来替代传统方法单一以共被引频次表征作者关联强度的方式。

3.1 引文全文文本挖掘及抽取

SPACA 的实施需要的不仅是著录信息,还有被引文献在施引文献全文中的所在引用句文本信息和位置信息,因此文本挖掘与抽取工作是必须的。

以 CNKI 数据库提供的 HTML 全文页面为例,通过 URL 将页面采集至本地。HTML 页面中包含有文献全文文本以及大量庞杂的数据标签,但半结构化的特征为数据抽取和内容分析提供了可能,通过编制解析器对 SPACA 所需的数据进行抽取和存储。见图 1。

通过 class_为"sup"的 <a> 标签,或 type_为"reference"的 <citation> 标签对引用句子实现定位(见图 1),对引文位置和引文内容文本信息的提取方法如下:

(a)引用位置信息的抽取:在 CNKI 现有的 HTML 格式全文中,主要以 <h3> 封装大标题, <h4> 封装小标题,在引用标签定位基础上,取父标签 <p> 并向前遍历以发现两种标签,直至获取 <h3> 标签为止,而 <h4> 可省缺,省缺时以 <h3> 代替。获取大标题和小标题后,就可在生成共被引对时,判断两引用句的位置是否接近,以辅助相关强度度量。

(b)引用句文本信息提取:本研究采用的基本方法为,由标签位置向前、后遍历,利用正则式进行判断,直到前邻接标签内容为不超出段落范围的完整文本,再利用句号分割文本,取尾部分作引用句前半段;后半段则同样以邻接标签文本的"。"为止作抽取范围,从而拼接成完整的内容文本。此外,研究针对多种引用标签格式、多种引用标签位置、小范围多个引用等情况都分别编制相应的引用句识别规则。

除以上数据以外,研究也在文后参考文献处对被引文献的基本信息进行了提取,包括作者、题名等,并筛除非期刊的条目。获得每篇参考文献的数据结构见图 2。

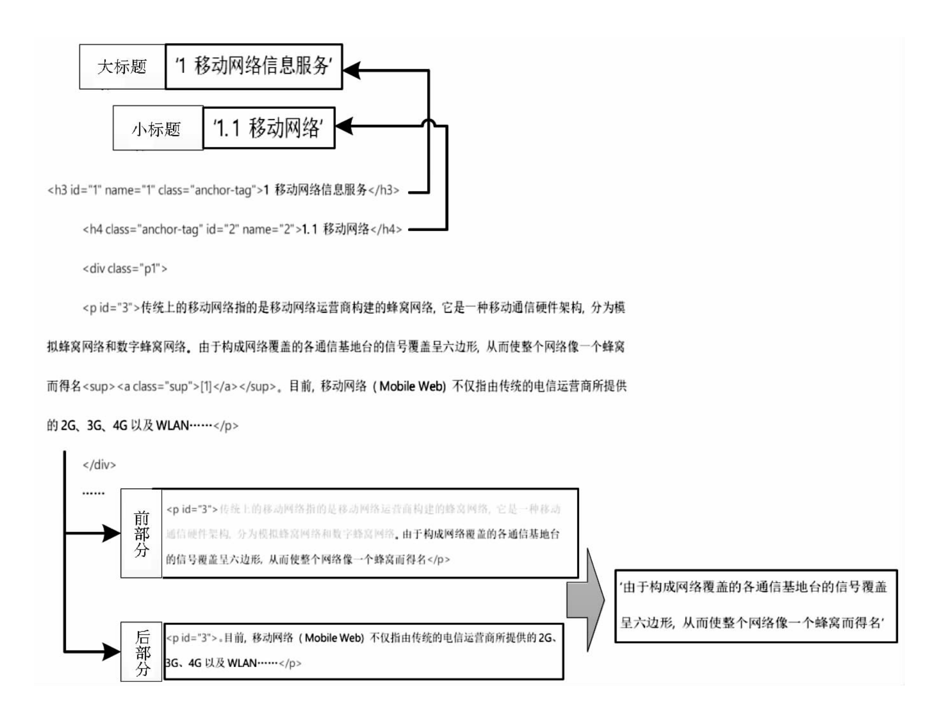
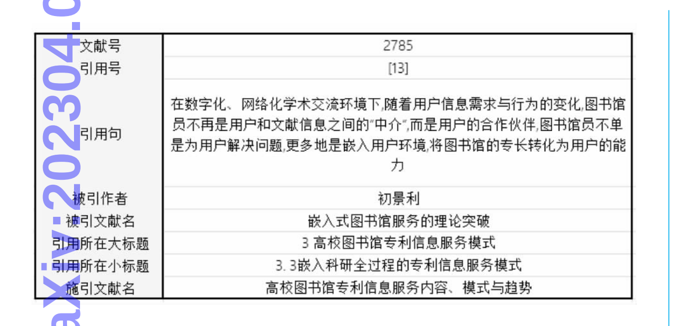


图 1　内容与位置数据抽取示意



文献号	2785
引用号	[13]
引用句	在数字化、网络化学术交流环境下,随着用户信息需求与行为的变化,图书馆员不再是用户和文献信息之间的"中介",而是用户的合作伙伴,图书馆员不单是为用户解决问题,更多地是嵌入用户环境,将图书馆的专长转化为用户的能力
被引作者	初景利
被引文献名	嵌入式图书馆服务的理论突破
引用所在大标题	3 高校图书馆专利信息服务模式
引用所在小标题	3.3嵌入科研全过程的专利信息服务模式
施引文献名	高校图书馆专利信息服务内容、模式与趋势

图 2　被引文献存储结构示意

3.2　基于领域语料的词嵌入模型训练

本文涉及引用句间的内容相似度计算,需要适当的自然语言处理工具提供词表示支持。以往涉及引用内容文本相似性计算的研究,多采用以固定词项为单位的词表示法,如直接利用独热编码(one-hot code)的词表示、词频－逆文档频率(TF-IDF)特征词抽取法等,这些方法虽简便,但词义和词形的割裂造成了两种问题:一是冗长的词维度带来的"维度灾难",二是无法适应现实中领域术语概念多样的表达方式。另一部分研究采用以 LDA、PLSA 为代表的语言概率主题模型,可建立起词、文档、主题或潜在语义联系,但在面对庞大的文本总量时,将产生昂贵的计算代价,这一问题对于常见的神经网络模型 Text RNN、Text CNN、BiLSTM 等也同样存在。上述缺点限制了以往改良方法的可用性和拓展性。

为使研究使用的词分布式表示更具语义内涵和概念性,并兼顾表示学习过程的效率与易拓展性,笔者选取 Word2Vec 词嵌入模型进行基于专业语料的预训练建模。Word2Vec 是一种从大量文本语料中以无监督的方式学习语义知识的浅层神经网络语言模型,其基本方法由 T. Mikolov 等提出,目前被大量应用于自然语言处理领域[40－41]。其包含 CBOW 和 Skip-gram 两种模式(见图 3),其基本原理是,对语料中某一词语与上下文窗口内词语间的联系进行建模,通过面向低维向量空间的映射,建立起每个词与相关词间的稠密向量联系,实现高效、高质量的词向量训练和优化[42]。相比基于词频的传统向量空间模型,Word2Vec 模型最大的特点是学习了词与词的发生语境联系。这一特性可以保证相似性度量的可靠性,使得词语不仅限于词语本身,而是与相关词语保持着主题上的相关,即语境甚至语义上的关联[43]。

Word2Vec 模型的质量建立在充分的语料训练基础上,这些大规模专业语料可来源于目标领域的全文文本,在经过基于领域文献关键词的定制化分词与预处理后用于模型训练,以学习领域词汇在语义与语境层面的权重向量。由于 Skip-gram 对低频词过于敏感,训练采用由窗口词共同对中心词预测并作共同权重调整的 CBOW 模式。训练形成模型效果如图 4 所示,对于该模型,当输入"学科化服务"这一词汇时,可获得与该词语义或语境相关性最大的部分词,如"学科服务""嵌入式学科服务""学科馆员""知识服务""学科

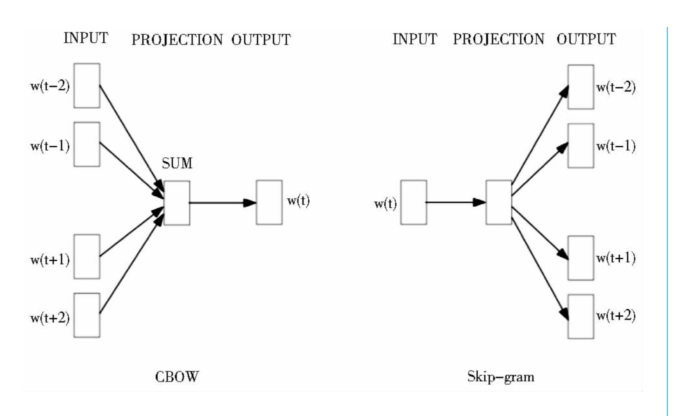


图 3　Word2Vec 两种不同模式的基本原理[40]

知识”等,这意味着即便撰写者在两个句子中应用了相同内涵不同形式的词汇,仍能通过模型准确判断两个句子的相似性。

```
In [260]: model.wv.most_similar('学科化服务')
Out[260]:
[('学科服务', 0.6190181374549866),
 ('学科馆员', 0.5274513363838196),
 ('嵌入式学科服务', 0.460089027881622 3),
 ('知识服务', 0.4552845656871795 7),
 ('信息服务', 0.4238835871219635),
 ('服务', 0.4038962125778198),
 ('知识化', 0.3913159370422363 3),
 ('嵌入式', 0.3877637386322021 5),
 ('学科知识', 0.383777078151703),
 ('教学科研', 0.3796292841434479)]
```

图 4　Word2Vec 模型效果示意

3.3　作者共被引关系强度算法

SPACA 作者相关强度算法综合考虑了共被引作者 A 与 B 的每一对共被引文献 $a_M(M=1,2,\cdots,n)$ 和 $b_N(N=1,2,\cdots,n)$ 的内容相似度 Content_similarity (a_M,b_N) 和被引位置相似度权重 P_Weight (a_M,b_N)(见式 1),并取 A 与 B 在共被引文献集中产生的最大相似度作为两者的相关强度(见式 2),以表征作者间可能产生的最大相关联系。

$$Similarity(a_M,b_N)=P_Weight(a_M,b_N)\cdot Content_similarity(a_M,b_N) \quad \text{式(1)}$$

$$Relevance(A,B)=\max\{Similarity(a_M,b_N)\} \quad \text{式(2)}$$

①在内容相似度 Content_similarity (a_M,b_N) 的计算中,利用了预训练获得的词嵌入向量和余弦相似度算法。基本原理是对 a_M 与 b_N 所在引用句 x 与 y 内分别包含的各词汇进行所含 i 维权重向量的叠加,构成句子向量 $\vec{W}_x$ 和 $\vec{W}_y$(见式 3),并作夹角余弦值计算,以量化测算内容相似度(见式 4)。此外,当内容相似度过低(小于 0.2)时,将舍弃此相似度值(置 0),以排除无联系的引用句对的干扰。

$$\vec{W}_{sentence}=\sum_1^n word_n(\vec{w}_1,\vec{w}_2,\cdots,\vec{w}_i)$$

$$\text{其中 } sentence=\{word_1,\cdots,word_n\} \quad \text{式(3)}$$

$$Content_similarity(a_M,b_N)=\cos(\vec{W}_x,\vec{W}_y)=\frac{\sum_i \vec{W}_x\vec{W}_y}{\sqrt{\sum_i \vec{W}_x^2}\sqrt{\sum_i \vec{W}_y^2}} \quad \text{式(4)}$$

②在被引位置相似度权重 P_Weight (a_M,b_N) 的计算中,采取以下算法:当两处被引发生于同一章节下时权重系数为 p,若进而发生于同一小节下则再乘权重系数 q(式 5),若均不同则权重为 1,进而利用位置权重对内容相似度进行加权。

$$P_Weight(a_M,b_N)=pos(x,y)=\begin{cases}1,(P_{chap.}(x)\neq P_{chap.}(y)\ \&\ P_{sec.}(x)\neq P_{sec.}(y))\\ p,(P_{chap.}(x)=P_{chap.}(y)\ \&\ P_{sec.}(x)\neq P_{sec.}(y))\\ p\cdot q,(P_{chap.}(x)=P_{chap.}(y)\ \&\ P_{sec.}(x)=P_{sec.}(y))\end{cases} \quad \text{式(5)}$$

为对 p 与 q 的取值进行调优,笔者进行了引用位置与内容相似度的相关性探索:将实验全文数据中的 177 617 条共被引对数据按临近度划分为文章级、章节级、小节级类型(TYPE),并对每对共被引对的内容相似度进行基于词嵌入模型的计算,获得相似度值(SIM)。由于 SIM 值不属正态分布,但具有方差齐性,故采用 Welch-Anova 和非参数检验法 Kruskal-wallis-Anova,并进行多重比较。结果显示,各组数据间分布相同的显著性远小于 0.05(见图 5),而非参数检验 $P<0.05$(见图 6),即不同位置级别的共被引对在 SIM 值分布上存在差异。

由表 1 可知,当邻近度由文章级提升至大章节级

Multiple Comparisons

Dependent Variable SIM
Test　Games-Howell

(I) TYPE	(J) TYPE	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval Lower Bound	95% Confidence Interval Upper Bound
1.0	2.0	-.041134*	.001155	.000	-.04384	-.03843
	3.0	-.104503*	.001634	.000	-.10833	-.10067
2.0	1.0	.041134*	.001155	.000	.03843	.04384
	3.0	-.063369*	.001676	.000	-.06730	-.05944
3.0	1.0	.104503*	.001634	.000	.10067	.10833
	2.0	.063369*	.001676	5.1224E-9	.05944	.06730

*. The mean difference is significant at the 0.05 level.

5.1224E-9

图 5　Welch-Anova 多重比较结果

和小章节级,中低相似(<0.5)的共被引对数量占比由 48.80% 降低到了 42.64% 和 31.816%,文本平均相似度累积提高了 8.48% 和 21.55%, 即邻近度每提升

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
1	The distribution of SIM is the same across categories of TYPE.	Independent-Samples Kruskal-Wallis Test	.000	Reject the null hypothesis.

Asymptotic significances are displayed. The significance level is .05.

图 6　非参数检验结果

一级,平均相似度约提升为上一级的 1.1 倍。综上,可以得出结论:3 种位置级别内的共被引对的相似度存在差异,处于逻辑结构上更小范围的共被引对,有更少内容不相关的可能,而倾向于更高的内容相似性。参考位置邻近度提升带来的相似度变化倍率,可将参数 p 与 q 均近似设为 1.1。

表 1　三组数据基本情况表

位置类型	数量	标准偏差	中位数	中低相似数量占比(<0.5)	平均相似度	平均相似度累积提升比	平均相似度变化倍率
文章级	85 166	0.225	0.507	48.80%	0.485	0.00%	1
大章节级	68 108	0.224	0.544	42.64%	0.526	8.48%	1.085
小节级	24 343	0.225	0.611	31.82%	0.589	21.55%	1.120
Total	177 617	0.228	0.536	44.11%	0.515	-	-

3.4　SPACA 方法的特点

与相关研究比较,笔者提出的 SPACA 方法具以下两个主要特点:一是 SPACA 方法利用基于领域语料的 Word2Vec 浅层神经网络模型建立了领域词间的语义及语境关联,以此计算引用句间的内容相似度,而不是基于作者共现与否、固定词项或小范围内的主题概率,保证作者间联系强度的可靠性以及该方法在复杂领域内的应变性;二是综合利用了引用发生的位置和主题两种全文信息,以加权的形式融合,综合表征作者间的关系强度,使关系强度计算时考虑要素更为多元,保证了方法投入应用时的稳定性。

4　实证研究

为进行 SPACA 的效果实证,探究其在复杂学科中的可应用性,笔者设计了对比实验,以比较 SPACA 与传统方法的效果差异,这种效果主要体现在:①对作者间联系揭示的准确度;②对学科中领域结构的识别程度。

面向这一目标,笔者将实验领域设定为国内图情学科,这出于两点考虑:其一,图情学科作为交叉性很强的学科,学者分布复杂,选定这类难度较大学科更易于显现两种方法分析效果的不同;其二,图情学科不同于基础科学学科,其专业术语缺乏如 MeSH 等词表限定,具多面性、多变性的特点,在国内环境下更是如此,在这样的领域进行对比实验,更能证明这一方法的可应用性和可拓展性。

4.1　实验流程概览

流程框架见图 7。首先利用采集器从数据源收集在线全文页面的 URL,并下载 HTML 页面至本地数据库。利用解析器提取全文文本信息、被引文献基本信息、引用句文本和位置信息,将相应信息分别传送给 ACA 和 SPACA。在对作者相关强度进行不同的计算后,分别构建作者共被引矩阵。在此基础上,利用网络分析和因子分析进行矩阵数据的直观呈现,最后对产生结果进行对比和讨论。

4.2　实验方法及过程

4.2.1　实验数据来源

实验数据来源为 CNKI 中文期刊数据库,初始时间窗设定为国内图情领域近 10 年的研究。在范围选定上,使用该库的文献分类目录,将主题范围框定在信息科技下的"图书情报与数字图书馆",时间范围定位于近 10 年(2009 -2019 年),将期刊级别限定在"SCI 来源期刊""EI 来源期刊" "核心期刊""CSSCI""CSCD"几类提供的核心目录中。仅选取每年被引量排行前 500 名的文献,这是因为高被引文献的质量相对较高,对于被引作者价值的体现也更具说服力。其中 2019 年由于距今时间较近,仅取前 142 篇(被引在 2 次以上)。

在充分遵循 CNKI 库访问规则与负载量的基础上,利用数据采集工具和 Python 编制网页采集规则,对 2009 -2019 年间图情领域核心期刊中的高被引论文进行在线全文 HTML 页面的 URL 采集,计划采集 5 142 条,排除不支持 HTML 的文献(占比约 9%),共采集条目 4 664 条(见图 8),通过脚本将 HTML 页面下载至本地。

4.2.2　数据抽取及共被引矩阵构建

利用脚本对 HTML 页面解析后,提取每条被引文献的基本著录信息、引用句文本信息、引用位置信息(参见 3.1 节),最终获得共计 23 572 条被引文献条目(见图 9)。其中,对 ACA 仅提供基本著录信息,如引用号、被引作者、被引文献名;而对 SPACA 提供所有信息,并以文献库中所有文本语料训练而成的 Word2Vec 模型提供语义相关支持(参见 3.2 节)。

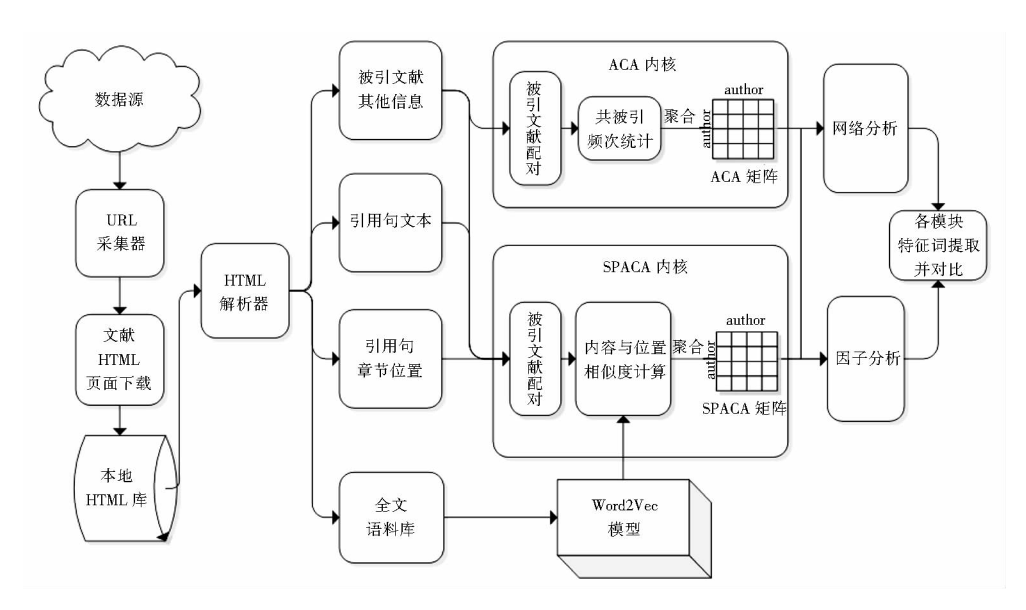


图 7　对比实证总流程

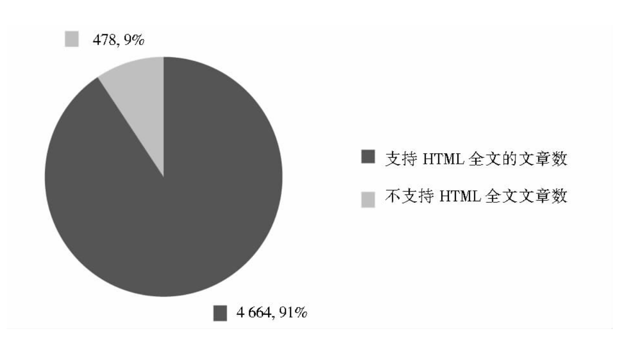


图 8　本研究采集数据中支持 HTML 格式的文献占比

9507	2781	[14]	图书馆对鉴别、开发、协调、提	['程宏利']	美国高校图书馆为远程教育服务的	4	面对MOOC大潮	4.3开发课程研究	MOOC环境下对高校图书馆信息服务工作的思考
9508	2781	[15]	加强与教师的合作，将MOOC与自	['张振虹'，'刘文']	从OCW课堂到MOOC学堂：学习本源的	4	面对MOOC大潮	4.3开发课程研究	MOOC环境下对高校图书馆信息服务工作的思考
9509	2783	[8]	而且在调查中发现，更大的空间	['闫小芬']	新时期美国大学图书馆馆舍建设趋	4	结论	4.3 优化空间资	在"纸张崇拜"与"数字拥戴"之间——高校
9510	2785	[2]	但是随着用户群体所处信息环境	['赵慧清']	高校图书馆专利咨询服务现状调研	0	研究简述	0 研究简述	高校图书馆专利信息服务内容、模式与趋势
9511	2785	[3]	图书馆对专利服务缺乏重视，处	['杨丽']	高校专利信息服务调查分析	0	研究简述	0 研究简述	高校图书馆专利信息服务内容、模式与趋势
9512	2785	[4]	曹湘博等构建了横向上包括研究	['曹湘博'，'曹锦丹']	面向技术创新的图书馆专利服务模	1	研究综述	1 研究综述	高校图书馆专利信息服务内容、模式与趋势
9513	2785	[5]	张簧玥等构建了横向上包括研	['张簧玥'，'肖国华']	高校图书馆专利信息服务研究	1	研究综述	1 研究综述	高校图书馆专利信息服务内容、模式与趋势
9514	2785	[6]	邓菲等的研究关注了嵌入科研过	['邓菲'，'李宏']	嵌入科研的情报研究服务新模式探	1	研究综述	1 研究综述	高校图书馆专利信息服务内容、模式与趋势
9515	2785	[7]	辛小萍等研究了嵌入科研过程	['辛小萍'，'田晓阳'，'周钊']	嵌入科研过程的服务模式探讨——	1	研究综述	1 研究综述	高校图书馆专利信息服务内容、模式与趋势
9516	2785	[8]	吴鸣等研究了专利技术分析在	['吴鸣'，'王丽']	嵌入式学科情报服务实践——以支	1	研究综述	1 研究综述	高校图书馆专利信息服务内容、模式与趋势
9517	2785	[9]	邓仲华从服务主体、服务客体、	['邓仲华'，'李立睿'，'陆颖隽']	大数据环境下嵌入科研过程的信息	1	研究综述	1 研究综述	高校图书馆专利信息服务内容、模式与趋势
9518	2785	[10]	专利信息素养教育旨在增强用户	['王欣'，'何立民'，'池晓波']	"卓越计划"唤醒工程类学生专利	2	高校图书馆专	2.1第一层：专	高校图书馆专利信息服务内容、模式与趋势
9519	2785	[11]	早在2007年工程硕士教育指导委	['徐升权']	全日制工程硕士"知识产权"课程	2	高校图书馆专	2.1第一层：专	高校图书馆专利信息服务内容、模式与趋势

图 9　被引文献存储库示意

基于该文献库，在排查重名作者后，面向第一作者，为 ACA 和 SPACA 分别构建作者共被引矩阵，对于 ACA 矩阵，作者相关强度取决于两者在文献集中的总共现次数；对于 SPACA，以基于语义和位置相似的共被引文献相似度的最大值作为作者相关强度。最后，形成共 10 684 个作者节点，132 267 对 ACA 共被引作者对和 118 388 对 SPACA 共被引作者对。

4.2.3　网络分析

利用可视化网络分析有助于直观了解学科内作者的分布情况、学科知识结构，比较两种 ACA 方法的效果差异。选用 Gephi 为工具，以作者作为节点（node），以作者间的相关联系为边（edge），以相关强度计算结果为边权（edge weight）。

在网络分析中，将参与分析的节点的被引频次阈值设置为 2。这是因为低频引用（被引频次 <1）节点在过去 10 年仅被利用 1 次，无法保证该次利用的可靠性，且它们的数量占比很高（为 65.26%），达 6 973 个节点（见图 10），对网络分析干扰较大，故剔除。

为进一步探测作者共被引网络中的分布规律与聚类情况，笔者利用 Louvain 模块化算法[44]。这一算法的基本思想是，对网络中的每一点都利用边权进行与相近节点的聚类，并多次迭代直到网络模块度（modu-

larity)不再提升为止,对于大型网络而言,其具有高效、较高精度的特点。

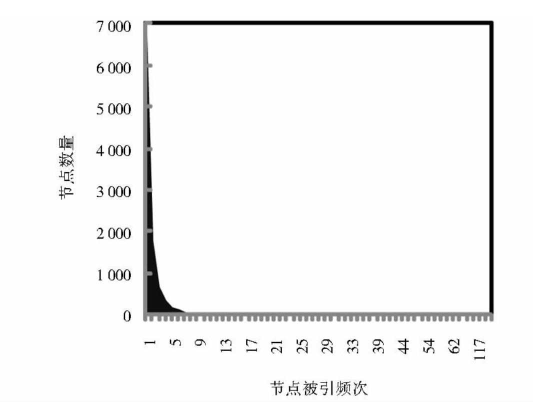


图 10　各被引频次节点数量分布

4.2.4　因子分析

在网络分析基础上,利用因子分析挖掘作者与领域、领域与领域间的相关性,并验证网络分析结果。因子分析是一种对潜在变量因子进行探测的降维方法,在以往相关研究中,其作为一种辅助聚类的工具被广泛采用。

笔者选取被引频次大于 15 次的 127 位作者为因子分析的对象。对于因子分析的输入数据,为降低过大的数值差距及过多零值对结果的干扰,研究将两种方法产生的作者共被引矩阵均转化为基于标准化欧几里得距离(euclidean distance)的相异度矩阵(dissimilarity matrix)。在观察两种方法的公因子总方差解释情况和特征值碎石图的基础上,抽取特征值(eigenvalues)大于 1 的公因子,鉴于各因子间存在低相关性,故选用直接斜交转轴法(direct oblimin rotation)对结果进行旋转,以获得更具可解释性的因子模式矩阵及载荷分布结果。

4.3　实验结果

4.3.1　网络分析结果

为使输出网络图更清晰可读,笔者以相同标准对两方法进行过滤。过滤规则包括:①在模块化后,过滤不成形的模块,即含节点占总数 1% 及以下或内部被引频次大于 15 次的核心节点数不足 2 个,以排除引用只发生于有限的小众研究主题或机构小范围内的情况,保证效果图的可读性。在此条件下,获得网络图中的主要模块:ACA 获得了 7 个模块,节点总数 3 005 个,占总数比为 80.98%,SPACA 获得了 12 个模块,节点数为 3 209 个,占节点总数比为 86.47%;②利用 K-brace 算法[45],对两端节点共同邻居节点不足 K = 10 的边进行剪枝,并调整边权重阈值,使网络输出简洁。

经上述过滤映射,形成图 11、图 12 所示的 ACA 输出网络图和 SPACA 输出网络图。其中,ACA 输出网络含 1 529 个节点和 3 156 条边,SPACA 输出网络含 1 502 个节点和 2 685 条边。初步观察网络图,在相同的布局算法 OpenOrd ＋ ForceAtlas2 的作用下,SPACA 的输出网络图(模块度 0.793)相比 ACA(模块度 0.697)在聚类分布上更集中,簇的辨识度更好。

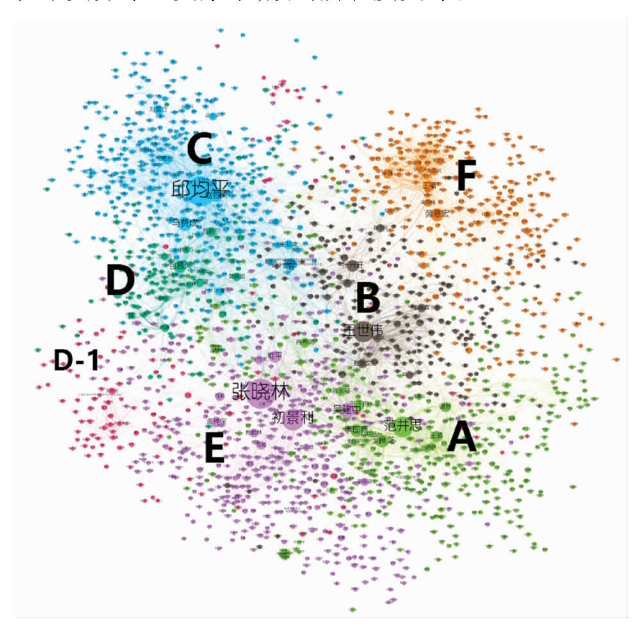


图 11　传统作者共被引分析－输出网络示意

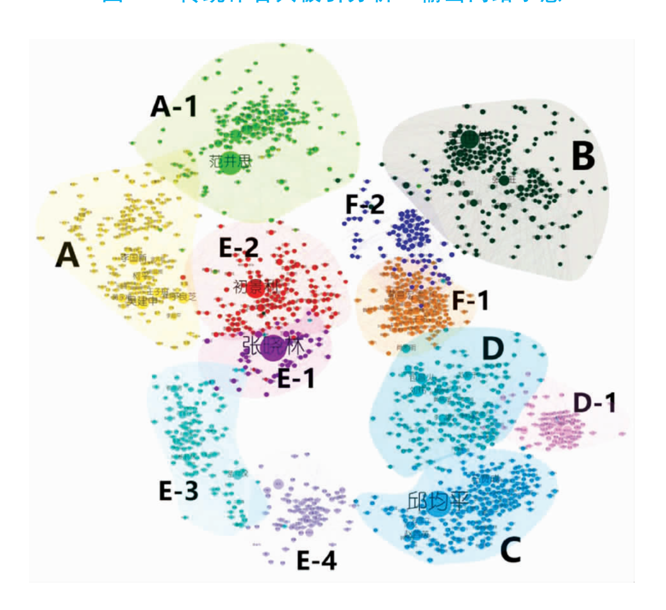


图 12　基于语义与位置相似的作者共被引分析——输出网络示意

为对各模块展开深入研究,对 ACA 和 SPACA 所产生的聚类结果进行特征提取。特征词提取工作流程围绕每一模块内两端节点均属该模块的边而展开,这些边中包含了被引文献对的标题与引用句。这种策略

的优点是能准确提取到模块形成过程中的支持性语料，使模块特征词提取更准确。特征词提取主要应用了 TF-IDF（TD）算法和 TextRank（TR）算法，并分别从特征值最高的 10 个词中筛选产生。由于引用句较短，LDA 与 LSA 方法的效果并不理想，故并未采用。

结合特征词提取结果和对模块内包含语料的人工检验，归纳出各模块的主题，并对两种方法中包含作者相似、特征词构成相似的模块标以相同 ID 号，形成了模块信息表（见表 2 和表 3）和主题对比表（见表 4）。需注意的是，此处的节点数和边数量是未进行过滤、剪枝的，即每一模块所真实含有的。

表 2　ACA 模块信息表（剪枝前）

编号	节点数量	边数量	特征词提取法	关键词	主题
A	610	527 2	TD	阅读推广 公共图书馆 活动 阅读疗法 文化	图书馆管理研究
			TR	阅读推广 公共图书馆 高校图书馆 阅读疗法 数字阅读	
D	227	159 9	TD	竞争情报 产业 情报 大数据 情报学	情报学理论与技术
			TR	竞争情报 情报 大数据 产业 情报学	
C	595	550 1	TD	网络 文献 领域 关键词 学科	科学计量与信息计量
			TR	知识图谱 关键词 文献 网络 可视化	
B	295	340 9	TD	智慧图书馆 智慧 技术 大数据 智慧服务	智慧图书馆及新技术应用
			TR	智慧图书馆 大数据 智慧服务 物联网 RFID	
E	713	572 8	TD	高校图书馆 资源 科研 创客空间 知识服务	研究图书馆数据与知识服务
			TR	高校图书馆 知识服务 学科服务 信息素养 数据管理	
F	383	398 0	TD	信息 情境 移动图书馆 用户 微信	移动图书馆与移动服务
			TR	情境 移动图书馆 信息 用户 微信	
D－1	182	162 1	TD	智库 服务 建设 高校图书馆 社科院	智库与智库研究
			TR	智库 高校图书馆 服务 建设 用户画像	

表 3　SPACA 模块信息表（剪枝前）

编号	节点数量	边数量	特征词提取法	关键词	主题
D	336	253 3	TD	竞争情报 情报学 关联数据 产业 大数据	情报学理论与技术
			TR	竞争情报 情报学 大数据 关联数据 分析	
A	399	266 2	TD	公共图书馆 建设 总分 创客空间 文化	图书馆管理研究
			TR	公共图书馆 创客空间 建设 文化 服务体系	
C	490	424 6	TD	网络 知识图谱 文献 领域 关键词	科学计量与信息计量
			TR	知识图谱 关键词 网络 文献 可视化	
E－3	252	161 0	TD	数据管理 高校图书馆 服务 数据素养 科研	研究型图书馆信息与数据素养、数据管理
			TR	高校图书馆 数据管理 数据素养 科研 信息素养	
B	389	390 9	TD	智慧图书馆 智慧服务 大数据 技术 用户	智慧图书馆与新技术应用
			TR	智慧图书馆 大数据 智慧服务 知识服务 物联网	
E－1	175	904	TD	知识服务 云计算 高校图书馆 资源 数字图书馆	知识服务与技术
			TR	知识服务 高校图书馆 云计算 数字图书馆 用户	
F－1	205	228 6	TD	情境 移动图书馆 用户 场景 模型	移动图书馆服务与技术
			TR	情境 移动图书馆 数字图书馆 服务 模型	
A－1	267	271 9	TD	阅读推广 阅读疗法 高校图书馆 数字阅读 儿童	阅读推广研究
			TR	阅读推广 高校图书馆 阅读疗法 活动 公共图书馆	
E－4	113	136 9	TD	开放存取 期刊 OA 机构知识库 质量	开放获取研究
			TR	开放存取 期刊 OA 机构知识库 模式	
E－2	303	198 1	TD	学科服务 用户 学科馆员 高校图书馆 信息素养	高校与研究图书馆学科服务
			TR	学科服务 高校图书馆 学科馆员 信息素养模式	
F－2	162	122 4	TD	微信 高校图书馆 传播 影响力 微信服务	新媒体行为与服务研究
			TR	微信 公众 高校图书馆 传播 阅读推广	
D－1	118	136 7	TD	智库 高校图书馆 社科院 竞争情报 决策	智库与智库研究
			TR	智库 高校图书馆 竞争情报 信息服务 决策	

表 4　ACA 与 SPACA 识别模块主题对比

ID	ACA	SPACA
A	图书馆管理研究（包括制度、阅读推广等）	图书馆管理（包括制度、服务等）
A－1		阅读推广研究
B	智慧图书馆与新技术应用	智慧图书馆与新技术应用
C	科学计量与信息计量	科学计量与信息计量
D	情报学与情报技术	情报学与情报技术
D－1	智库研究	智库研究
E	研究图书馆数据与知识服务	
E－1		知识服务与技术
E－2		研究图书馆学科服务
E－3		研究图书馆信息与数据素养研究
E－4		开放获取研究
F	移动图书馆与新媒体	
F－1		移动图书馆服务与技术
F－2		新媒体行为与服务研究

通过观察和对比两种方法所识别模块的网络输出图、模块关键词、主题情况、节点数量情况，笔者发现两种方法产生的一些模块是相似的，如以邱均平等为代表的 C 模块“科学计量与信息计量”、以王世伟等为代表的 B 模块“智慧图书馆与新技术应用”。此外，研究发现以下要点：

（1）一些在 ACA 中包含于同一模块的关键词，在 SPACA 中被划归至了两个或多个模块中。典型的例子包括 ACA 中的 F 模块“移动图书馆与移动服务”（383 个节点），其包含有的“移动图书馆”“微信”“公众”等词被分别划归入了 SPACA 的 F－1 移动图书馆服务与技术（205 个节点）和 F－2 新媒体行为与服务研究（162 个节点）模块下，联系网络输出图中的局部例证（如图 13 所示），发现作者的迁移也同样印证了这种细分的变化。经证实，如孔云、王保成等作者，相比于移动图书馆，更倾向于新媒体行为与服务的研究。

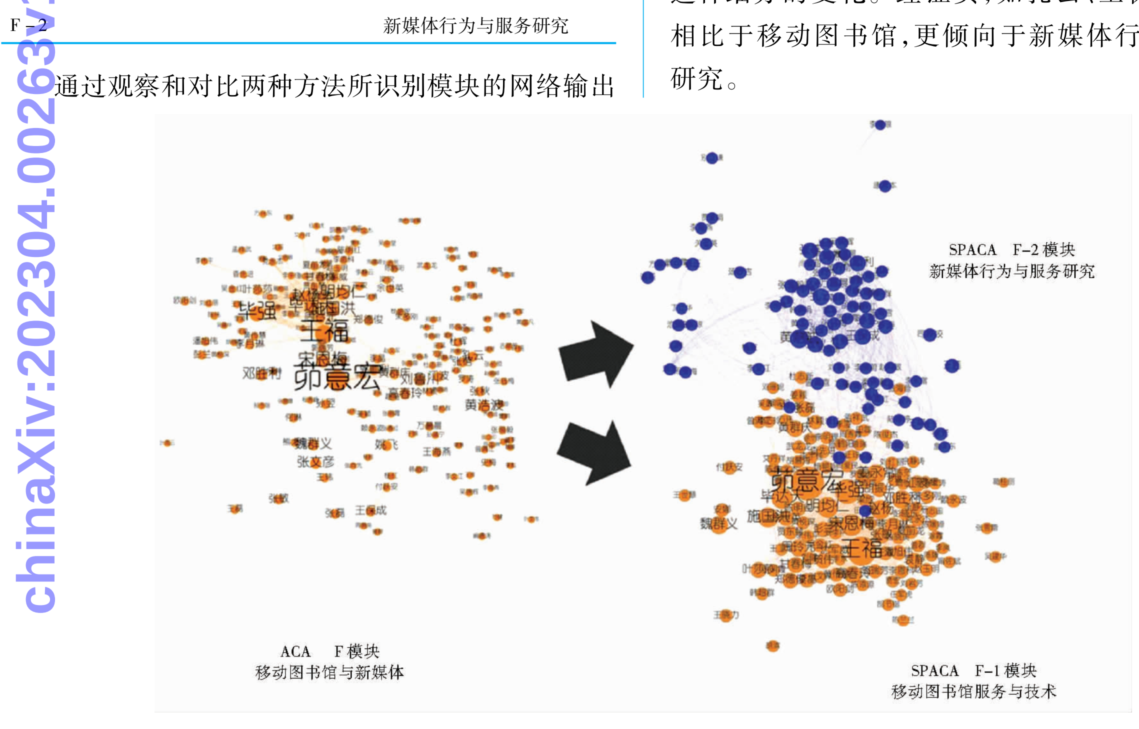


图 13　模块细分例证

SPACA 方法产生的上述细分情况，使这类较新、与某一领域关系密切却又自具特点的领域更易被发现，对这些发生细分变化的领域进行阴影标注，形成具有拓扑联系的网络输出图（如图 14 所示）。除 F 模块外，笔者还发现：①研究型图书馆服务下的各类型方向被细分：以张晓林为代表的 E－1“知识服务与技术”研究、初景利为代表的 E－2“学科化服务”研究、孟祥保和司莉为代表的 E－3“信息与数据素养、科研数据管理”研究、陈传夫等为代表的 E－4“开放获取”研究从 E 模块中被细分；②以范并思、王波为代表的 A－1 模块“阅读推广”，从关系极为密切的 A 模块“图书馆管理”这一涉及图书馆业务研究的主领域中被分离出来。以上现象均说明了，SPACA 能使共被引作者在共被引句的内容和位置层面，建立起更细粒度的联系。

（2）一些在 ACA 中识别不准确的节点，在 SPACA 中被矫正并移入了更恰当的作者群。典型例子包括：原属 E 模块“研究图书馆数据与知识服务”的作者柯平、吴建中，在 SPACA 中被归入了 A 模块“图书馆管理”，经查证，两位作者主要从事图书馆事业、图书馆评估、图书馆管理转型方面研究，更偏向于公共图书馆研究，这说明他们在 SPACA 中所归属的模块更为准确。在对相关的引用句对查考后发现，柯平、吴建中两位作

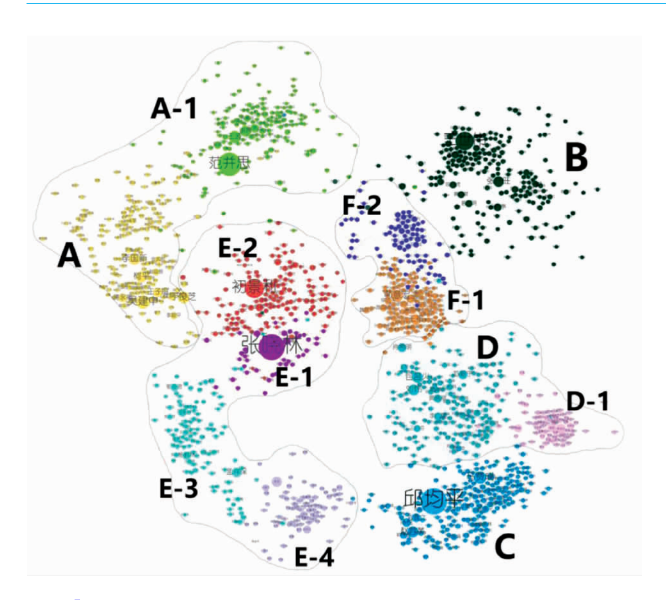


图 14 对细分结果阴影标注的 SPACA 输出网络

者与张晓林、初景利等作为知名学者,常在知识服务、开放获取等话题中因某些权威定义或理论而被共同引用,这使得这 4 位学者间共被引频次较高,最高共被引次数达 52 次,导致以频次为边权的 ACA 将这 4 位研究领域有一定差异的学者误归为同一模块(见图 15),造成了聚类误差。这在 SPACA 中得到了解决,柯平、吴建中所从事的研究领域与另两位作者存在内容上的差别,如图 16 所示,吴建中与初景利的边权重仅 6.30(为方便比较,边权重值统一放大 10 倍),柯平与初景利的边权重约为 8.4,而吴建中与柯平间的边权重达到了 11.89,这形成了吴、柯两位作者间的凝聚及与另两位作者间的分隔,此外,SPACA 甚至识别出了初、张两位作者间的不同。这说明考虑语义与位置信息的作者共被引分析更为准确,能纠正 ACA 所形成网络布局的不合理之处。

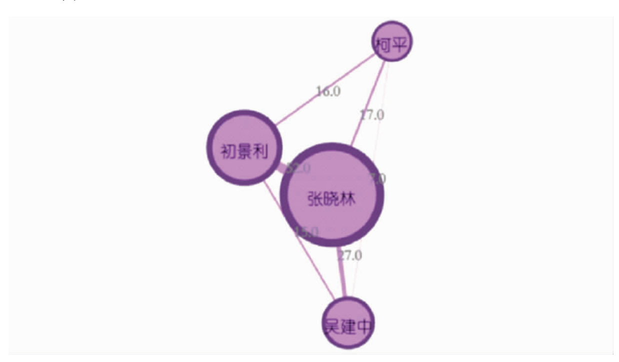


图 15 ACA 展现的作者联系示例

4.3.2 因子分析结果

在对 127 名主要作者进行的因子分析中,对于特征值大等于 1 的条件,ACA 提取到了 9 个公因子,总解释方差为 87.992%,SPACA 提取到了 13 个公因子,总解释总方差为 90.767%(见表 5)。SPACA 方法相较于 ACA 提取到了更多因子,且解释了稍多的方差,这初步说明 SPACA 的因子分析结果分布更为合理,效果优于 ACA。

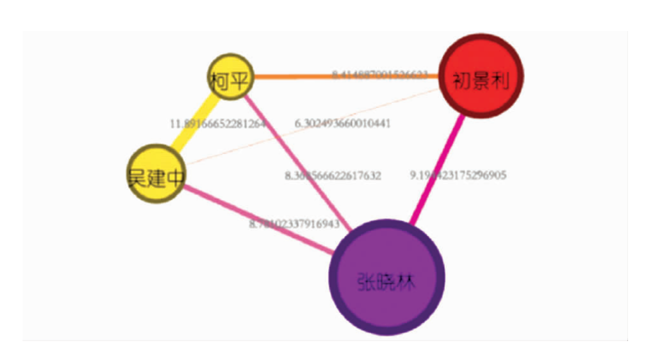


图 16 SPACA 展现的作者联系示例

表 5 因子分析概况

输入矩阵	公因子数量(Num of Factors)	解释总方差(Total variance explained)
ACA 相异矩阵	9	87.992%
SPACA 相异矩阵	13	90.767%

笔者进而对直接斜交旋转后产生的模式矩阵深入分析。首先对矩阵中各公因子(子领域)所含变量(作者)进行提取和统计,判断变量归属公因子的依据是载荷大于等于 0.3。因此,有部分变量可归属于多个公因子,也可不归属于任何识别出的公因子,这也与图情领域作者个人研究方向多元化的特点相似。其次,研究对每一公因子下的变量所代表的作者,与该公因子下其他作者的共被引关系(包含有引用句和引用标题)进行再提取,利用 TF-IDF 算法提取每个公因子的特征词,以客观地表征每一公因子所指代的领域。

研究发现,公因子所代表的主题与网络分析结果相似,ACA 和 SPACA 均能识别出包括 A-F 在内的学科一级领域,除部分因子间存在有主题重叠情况,少量主题未被识别外,因子分析与网络分析聚类结果能够实现一一对应,这也印证实验所用分析法的合理性。具体见表 6。

观察各公因子信息表(见表 6)及其包含作者,笔者有以下发现:

(1)ACA 识别出的公因子 F4 与 F5 中包含以王波、范并思、李国新、于良芝等为代表的 22 名作者,主要从事图书馆管理方面研究;这些作者在 SPACA 中被细分出了阅读推广这一子领域,王波、范并思等被归入阅读推广研究领域(F7)之下。ACA 中的 F1 与 F3 中包含以王世伟、储节旺、马晓亭、张兴旺等为代表的 39

表 6　因子分析结果

领域主题	ACA			SPACA		
	公因子	含作者数	最高载荷	公因子	含作者数	最高载荷
A 图书馆管理	F4,F5	22	0.924	F4	15	0.885
A－1 阅读推广				F7	10	0.922
B 智慧图书馆与新技术应用	F1,F3	39	1.053			
B－1 智慧图书馆理论				F2	14	0.939
B－2 新技术应用				F5	6	0.738
C 科学计量与信息计量	F7	16	1.016	F10	19	0.859
D 情报学与情报技术	F8	16	0.876	F6,F12	16	0.768
D－1 智库研究				F8	6	0.570
E 研究型图书馆学科服务、知识服务与开放获取	F6,F9	21	0.828	F1, F13	11	0.713
E－3 研究图书馆信息与数据素养、科研数据管理研究				F11	8	0.758
F 移动图书馆与新媒体	F2	10	1.019			
F－1 移动图书馆服务与技术				F3	20	0.883
F－2 新媒体行为与服务研究				F9	7	0.631

名作者,主要从事智慧图书馆与图书馆新技术应用方面工作;而在 SPACA 中,王世伟、储节旺等被归为智慧图书馆理论研究(F4)作者,而马晓亭、张兴旺等被归为新技术应用研究(F7)作者。

(2)ACA 识别出的 F7 与 SPACA 识别出的 F10 相似,都包含有以邱均平、赵蓉英等从事科学计量与信息计量的作者。该子领域由于术语、方法、引用对于整个领域来说相对稳定,因此在图情领域这一整体大视角下效果差异不大。

(3)ACA 中的 F8 包含李广建、包昌火等在内的情报与情报技术研究者 18 名,而 SPACA 除识别出了这一领域(F6、F12)外,还发现同属情报学研究范畴的智库研究领域(F8),代表作者包括黄如花、吴育良、李纲等。智库研究这一公因子最大载荷较低,内部变量在其他公因子均有分布载荷。这说明在图情领域内,智库研究是一个新兴子领域,目前其中的学者具有较强的研究领域交叉性。

(4)ACA 中的公因子 F6、F9 所包含有 21 名以张晓林、初景利为代表的从事研究型图书馆知识服务研究的作者,这对应于 SPACA 中的公因子 F1、F13。而在 ACA 的 F6 和 F9 下由于低载荷未被识别的杨鹤林、孟祥保等作者被分入信息素养和数据素养研究领域。与网络分析结果不同的是,E－1 知识服务、E－2 学科服务、E－4 开放获取未能被识别,仅被归统于 SPACA 中的 F1 和 F13 所表征的"研究型图书馆学科服务、知识服务与开放获取"大类中,这是由于这些领域划分较细,除部分知名作者外,其他作者被引频次不高,在 127 名核心作者中占比低而导致公因子特征值不足。

(5)ACA 中的公因子 F2 包含以毕达天、王福在内的从事移动图书馆与新媒体平台服务研究的作者。对于该领域,SPACA 识别出了移动图书馆服务与技术(F3)和新媒体行为与服务研究(F9)两个细分领域,前者以王福、毕强为代表,后者以黄浩波、高春玲为代表,这与网络分析结果类似。这是由于在图情领域的研究中,移动服务与社交媒体平台研究往往联系紧密,许多作者同时从事两方面的研究,因此传统方法难以区分两个子领域,但 SPACA 的自身特点却使这种细分成为可能。

5　讨论

综合实验结果,笔者得出以下结论:

(1)笔者所提出的 SPACA 方法实现了对作者间联系的更准确刻画。这表现为:网络分析中 SPACA 展现了更明晰的节点聚类、后期人工检验中更少的不准确分类作者节点,这很大程度上是因为 SPACA 计算中纳入了语义和位置信息,以此作为衡量作者联系的具丰富内涵的标度。这种标度更细化地展现了联系强度,深入内容语义层面,考虑了共被引文献间的联系及被共引作者间隐含的主题相关性。其有效克服了 ACA 相关性度量指标单一、存在作者联系强度与个人出现频次间的次带联系等天然缺陷[46],使得两作者会因被引内容的语义和位置的相似而聚集,而不受热门主题等带来的共现频次波动干扰,更符合现实情况。

(2)SPACA 在学科知识结构发现中实现了更细致的子领域识别。在本研究所设计的控制变量条件下,不论在网络分析还是因子分析检验中,相比传统方法,

SPACA 都识别出了更多的子领域。这一差距的产生，从宏观角度上看，是因为这些小型领域较新或尚处于融汇发展阶段，这些领域所含作者节点体量不大，在结构上与其他领域联系密切，在主题上也与其他领域交叉密切，构成了传统方法的“盲点”。从微观角度上看，是因为研究者们往往倾向于在文章前段位置引用领域权威学者的知名论断作研究铺垫（见图 17），这导致多数形成的共被引对集合在实质上是一种基于施引者个体意图的弱连接集，其本质是间接的，强度范围是狭窄而模糊的，而真正相关的连接却极易被忽略。而 SPACA 方法在考虑语义、语境及位置相关性的深层层面建立起集合内的作者间的强连接，剔除无关的弱连接，使得一些新兴的、对其他领域具依赖性的小型领域也能被“剥离”出来，在对领域了解或有专家辅助判断的条件下，甚至能够发现学科内的拓扑结构。

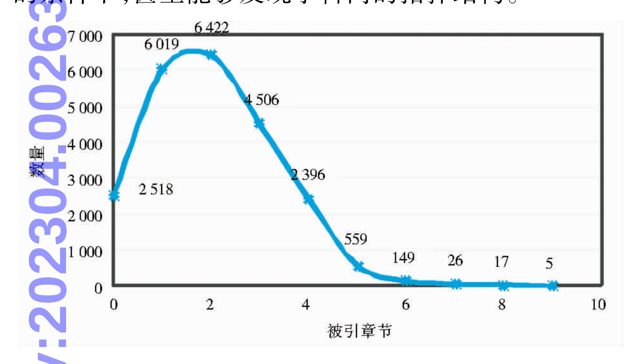


图 17　引用发生的章节分布

（3）SPACA 方法在面向大型交叉学科时具有可应用性与良好效果。研究中网络分析与因子分析结果的一致性和相互印证性，证明 Word2Vec 词嵌入模型可有效应用于社会科学领域和交叉学科领域文本间的内容相似度判别，这是因为其构造了概念词汇间的语境，进而语义上的联系，使得即便在术语复杂的领域仍具拓展性。同时，实验证明了所提出的研究流程能够克服国内数据库难用于获取全文数据进行引文内容分析的难点。

（4）本研究也存在着不足，具体如下：①部分文献暂未提供 HTML 页面，虽然只占 9%，但无法排除导致重要文献遗漏的可能；②由于所用全文文本库部分格式的不规范性，以及文献引用时的随意性，少部分引用句抽取存在异常；③模型训练语料在体积上仍然不够，此外，即便模型在训练前利用了领域关键词进行定制化分词，但仍有着被过度细分的词，这些问题使模型中存在着部分无用的“噪声词”，对结果产生了一定干扰；④面对从事多领域、多方向研究的学者，本文的文本挖掘和分析方法仍缺乏深度，如何实现更精准的 ACA 分析效果亟待进一步研究。

6　结语

笔者提出了一种改良型作者共被引分析法 SPACA，通过采集领域全文文本，利用 Word2Vec 词嵌入模型和位置加权相结合的方法计算共被引作者关联强度，并与传统基于共现频次的作者共被引分析法进行对比实验，实验证明 SPACA 能更准确地识别作者联系，发现更细致而富有立体感的学科子领域分布，同时，也为充分利用国内引文全文进行文献研究提供了可参考路径。

笔者认为，未来的研究应尝试从以下方面拓展：①对引文的内容语义和位置信息做更深层的挖掘，如纳入本体与概念图等方法来表征更深层的语义；②纳入更多有意义的指标数据到分析中来（如强度、动机、情感），并在细节、参数调整、聚类方法等方面作更多的优化；③思考引文内容分析在其他信息计量方法上进行应用和创新的可能。当下，大学科环境正呈现着快速演化、交叉相融等复杂态势，图情学界应对引文分析方法本身的机理作更多的思考与探讨，由仅仅依赖于表层著录信息，转向对引文全文本的充分挖掘和利用。

致谢：感谢中国科学院文献情报中心袁军鹏研究员，成都文献情报中心杨志萍研究员、陈云伟研究员，福建师范大学图书馆学系傅文奇教授，武汉大学信息与管理学院余凡博士，三位外审专家及《图书情报工作》编辑部为本研究提出的宝贵建议。

参考文献：

[1] WHITE H D, GRIFFITH B C. Author cocitation: a literature measure of intellectual structure. [J]Journal of the American society for information science, 1981, 32(3):163－171.

[2] BAYER A E, SMART J C, MCLAUGHLIN G W. Mapping intellectual structure of a scientific subfield through author cocitations [J]. Journal of the American society for information science, 1990, 41(6):444－452.

[3] BOYACK K W, SMALL H, KLAVANS R. Improving the accuracy of co-citation clustering using full text[J]. Journal of the American society for information science and technology, 2013,64(9):1759－1767.

[4] DING Y,ZHANG G,CHAMBERS T. Content-based citation analysis: the next generation of citation analysis[J]. Journal of the association for information science and technology, 2014, 65(9):1820－1833.

[5] 胡志刚. 全文引文分析方法与应用[D]. 大连:大连理工大学, 2014.

[6] 刘盛博,丁堃,唐德龙. 引用内容分析的理论与方法[J]. 情报理论与实践,2015,38(10):27 - 32.

[7] LIU S, CHEN C. The differences between latent topics in abstracts and citation contexts of citing papers[J]. Journal of the American society for information science and technology, 2013, 64(3):627 - 639.

[8] DING Y, SONG M, HAN J, et al. Entitymetrics: measuring the impact of entities[J]. PLoS ONE, 2013, 8(8),1 - 14.

[9] 章成志, 徐庶睿, 卢超. 利用引文内容监测多学科交叉现象的方法与实证[J]. 图书情报工作, 2016, 60(19): 108 - 115.

[10] NANBA H, OKUMURA M. Towards multi-paper summarization using reference information[C]//The commitee of international joint conferences on artificial intelligence. Proceedings of the 16th international joint conferences on artificial intelligence. San Francisco: Morgan Kaufmann Publishers, 1999: 926 - 931.

[11] TEUFEI S, SIDDHARTHAN A, TIDHAR D. An annotation scheme for citation function[C]//ALEXANDERSSON J. Proceedings of the 7th SIGdial workshop on discourse and dialogue. Stroudsburg: Association for Computational Linguistics,2009:80 - 87.

[12] ZAFAR L, AHMED U, ISLAM M A. Citation context analysis using word-graph[C]//IEEE 2019 2nd International conference on communication, computing and digital systems (C-CODE). Islamabad: the Institute of Electrical and Electronics Engineers,2019: 120 - 125.

[13] ABU-JBARA A, EZRA J, RADEV D R. Purpose and polarity of citation: towards NLP-based bibliometrics[C]//Proceedings of the 2013 conference of the North American chapter of the association for computational linguistics: human language technologies. Atlanta: Association for Computational Linguistics, 2013: 596 - 606.

[14] 雷声伟,陈海华,黄永,等. 学术文献引文上下文自动识别研究[J]. 图书情报工作, 2016, 60(17): 78 - 87.

[15] ANGROSH M A, CRANEFIELD S, STANGER N. Context identification of sentences in related work sections using a conditional random field: towards intelligent digital libraries[C]// HUNTER J. Proceedings of the 10th annual joint conference on digital libraries. New York: ACM, 2010: 293 - 302.

[16] ZHU X, TURNEY P, LEMIRE D, et al. Measuring academic influence: not all citations are equal[J]. Journal of the association for information science and technology, 2015, 66(2): 408 - 427.

[17] SOMBATSOMPOP N, KOSITCHAIYONG A, MARKPIN T, et al. Scientific evaluations of citation quality of international research articles in the SCI database: thailand case study[J]. Scientometrics, 2006,66(3):521 - 535.

[18] CHEN C, LIU Z. Where are citations located in the body of scientific articles? A study of the distributions of citation locations[J]. Journal of informetrics,2013,7(4):887 - 896.

[19] LU C, DING Y, ZHANG C. Understanding the impact change of a highly cited article: a content-based citation analysis[J]. Scientometrics, 2017, 112(2), 927 - 945.

[20] DING Y, LIU X, GUO C, et al. The distribution of references across texts: some implications for citation analysis[J]. Journal of Informetrics, 2013, 7(3):583 - 592.

[21] HOU W R, LI M, NIU D K. Counting citations in texts rather than reference lists to improve the accuracy of assessing scientific contribution[J]. Bioessays,2011,33(10):724 - 727.

[22] ELKISS A, SHEN S, FADER A, et al. Blind men and elephants: what do citation summaries tell us about a research article? [J]. Journal of the American society for information science and technology, 2008, 59(1): 51 - 62.

[23] CALLAHAN A, HOCKEMA S, EYSENBACH G. Contextual cocitation: augmenting cocitation analysis and its applications[J]. Journal of the American society for information science and technology, 2010, 61(6): 1130 - 1143.

[24] GIPP B, BEEL J. Citation proximity analysis (CPA)-a new approach for identifying related work based on co-citation analysis [C]// LARSEN B, LETA J. Proceedings of ISSI 2009 - The 12th international conference on scientometrics and informetrics. Rio de Janeiro: BIREME/PAHO/WHO and Federal University of Rio de Janeiro, 2009: 571 - 575.

[25] GIPP B. Citation proximity analysis-a measure to identify related work[D]. Magdeburg: Otto-von-Guericke University, 2006.

[26] LIU S, CHEN C. The effects of cocitation proximity of cocitation analysis[C]//NOYONS E, NGULUBE P, LETA J. Proceedings of ISSI 2011 - the 13th international conference on scientometrics and informetrics. Durban: Leiden University and University of Zululand, 2011:474 - 484.

[27] AN J, KIM N, KAN M Y, et al. Exploring characteristics of highly cited authors according to citation location and content[J]. Journal of the Association for information science and technology, 2017, 68(8):1975 - 1988.

[28] ETO M. Evaluations of context-based co-citation searching[J]. Scientometrics, 2013, 94(2):651 - 673.

[29] 赵蓉英, 郭凤娇, 曾宪琴. 基于位置的共被引分析实证研究[J]. 情报学报, 2016, 35(5):492 - 500.

[30] JEONG Y K, SONG M, DING Y. Content-based author co-citation analysis[J]. Journal of informetrics, 2014, 8(1): 197 - 211.

[31] LU K, WOLFRAM D. Measuring author research relatedness: a comparison of word-based, topic-based, and author cocitation approaches[J]. Journal of the American society for information science, 2012, 63(10): 1973 - 1986.

[32] 祝清松, 冷伏海. 基于引文内容分析的高被引论文主题识别研究[J]. 中国图书馆学报, 2014,40(1): 39 - 49.

[33] 刘盛博, 张春博, 丁堃, 等. 基于引用内容与位置的共被引分析改进研究[J]. 情报学报, 2013, 32(12): 1248 - 1256.

[34] KIM H J, JEONG Y K, SONG M. Content- and proximity-based author co-citation analysis using citation sentences[J]. Journal of informetrics, 2016, 10(4): 954 - 966.

[35] 李秀霞,邵作运. 融入内容信息的作者共被引分析——以学科服务研究主题为例[J]. 图书情报工作,2016,60(1):98 - 104,141.
[36] 肖雪,陈云伟,邓勇. 基于节点内容及拓扑结构的引文网络社团划分[J]. 图书情报知识,2017(1):89 - 97.
[37] 张艺蔓,马秀峰,程结晶. 融合引文内容和全文本引文分析的知识流动研究[J]. 情报杂志,2015, 34(11): 50 - 54,49.
[38] DING Y, ZHANG G, CHAMBERS T, et al. Content-based citation analysis: the next generation of citation analysis[J]. Journal of the association for information science and technology, 2014, 65(9): 1820 - 1833.
[39]赵蓉英,曾宪琴,陈必坤. 全文本引文分析——引文分析的新发展[J]. 图书情报工作,2014,58(9):129 - 135.
[40] MIKOLOV T, CHEN K, CORRADO G, et al. Efficient estimation of word representations in vector space[EB/OL]. [2019 - 12 - 12]. https://arxiv. org/abs/1301. 3781.
[41] MIKOLOV T, SUTSKEVER I, CHEN K, et al. Distributed representations of words and phrases and their compositionality[C]// BURGES C. Advances in neural information processing systems. Lake Tahoe: Neural Information Processing Systems Foundation, 2013: 3136 - 3144.
[42] 唐晓波,翟夏普. 基于本体和 Word2Vec 的文本知识片段语义标引[J]. 情报科学,2019,37(4):97 - 102.
[43] LAW J, ZHUO H H, HE J, et al. LTSG: Latent topical skip-gram for mutually improving topic model and vector representations [C]// LAI J H. Pattern recognition and computer vision PRCV 2018. Cham: Springer, 2018:375 - 387.
[44] BLONDEL V D, GUILLAUME J, LAMBIOTTE R, et al. Fast unfolding of communities in large networks[J]. Journal of statistical mechanics: theory and experiment, 2008(10): P10008.
[45] UGANDER J, BACKSTROM L, MARLOW C, et al. Structural diversity in social contagion[C]//GRAHAM R L, JOLLA L. Proceedings of the national academy of sciences of the United States of America. Washington: PNAS, 2012: 5962 - 5966.
[46] 苑彬成,方曙,刘合艳. 作者共被引分析方法进展研究[J]. 图书情报工作,2009,53(22):80 - 84.

Empirical Study of a Semantic and Proximity-based Author Co-citation Analysis Method

Zhang Ruhao
Chengdu Library and Information Center, Chinese Academy of Sciences, Chengdu 610041
Department of Library and Information Science, School of Economics and Management,
University of Chinese Academy of Sciences, Beijing 100049
National Science Library, Chinese Academy of Sciences, Beijing 100190

Abstract: [**Purpose/significance**] The author co-citation analysis is an vital method to explore the domain knowledge structure. In the context of complex development of disciplines, the author's relevance measure based on the co-citation frequency is quite controversial. The study proposed an improved method for author co-citation analysis based on the similarity of content semantics and the proximity of locations. [**Method/process**] Based on the introduction of its basic principles, the field of LIS was used as an example to demonstrate the effect of the method, a full-text mining of citations for CNKI Chinese journals was conducted, and the citing sentences and reference positions were then extracted. Combined with pre-trained domain word embedding models, the deep correlation between the co-cited literature and the strength of the connection between the authors were measured. A network analysis and a factor analysis were then used to compare the differences on effects between the method and the traditional method. [**Result/conclusion**] The results show that the method can more accurately identify the correlation strength between authors, and find more detailed subject knowledge structure, and has a certain scalability and applicability.

Keywords: author co-citation analysis citation content analysis co-citation proximity analysis citation in full-text knowledge structure